CLOUD APPLICATION DEVELOPMENT (GROUP 1)

PHASE 2: ASSIGNMENT NOTEBOOK SUBMISSION

NAME : Tamil selvi.P

PROJECT TITLE : Big Data Analysis with IBM Cloud Databases[05]

EMAIL ID : tamiljesus2004@gmail.com
GUIDED BY: MRS.J.HEMALATHA

SPOC NAME : MR. P. VIGNESH

GIT HUB REPOSITORY URL : **https://github.com/TamilSelvip-2004/Bigdata-analysis/settings/keys**

BIG DATA ANALYSIS

Phase 2: Innovation

Consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

Random Forest:

Ensemble Learning: Random Forest is an ensemble of decision trees, which combines multiple models for improved accuracy.

Decision Trees: Each tree in the forest is constructed based on different subsets of data, reducing overfitting.

Importance Scores: It provides feature importance scores, helping you identify which features are most influential.

Gradient Boosting:

XGBoost, LightGBM, CatBoost: These are gradient boosting libraries known for their speed and performance.

Boosting Process: Boosting iteratively improves the model's performance by focusing on data points that were previously misclassified.

Regularization: These algorithms offer regularization techniques to prevent overfitting.

Neural Networks:

Convolutional Neural Networks (CNNs): Ideal for image recognition tasks, they use convolutional layers to detect patterns in images.

Recurrent Neural Networks (RNNs): Suitable for sequential data, they capture temporal dependencies.

Deep Learning: Deep neural networks with many layers can learn intricate patterns but require substantial data and computational resources.

Support Vector Machines (SVM):

Margin Maximization: SVM aims to find the hyperplane that maximizes the margin between classes.

Kernel Trick: Non-linear SVMs use kernel functions to transform data into higher-dimensional space for better separation.

Clustering Algorithms:

K-Means: Divides data into clusters based on centroids, often used for customer segmentation.

DBSCAN: Density-based clustering identifies clusters of varying shapes and sizes.

Hierarchical Clustering: Creates a tree-like hierarchy of clusters, useful for understanding data structure.

Time Series Analysis:

ARIMA (AutoRegressive Integrated Moving Average): Suitable for stationary time series data.

Prophet: Developed by Facebook, it handles time series with seasonality and holidays.

LSTM (Long Short-Term Memory): A type of RNN, effective for sequential data with long-range dependencies.

Dimensionality Reduction:

Principal Component Analysis (PCA): Reduces data dimensions while preserving as much variance as possible.

t-SNE (t-distributed Stochastic Neighbor Embedding): Useful for visualizing high-dimensional data in lower dimensions.

Ensemble Methods:

Stacking: Combines predictions from multiple models with a meta-model for improved accuracy.

Bagging: Bootstrap aggregating averages predictions from multiple models to reduce variance.

Conclusion:

- Incorporating these techniques involves careful data preprocessing, such as handling missing values, scaling, and encoding categorical variables. Additionally, hyperparameter tuning and cross-validation are essential to optimize model performance.
- The choice of algorithm depends on the nature of your data and the specific objectives of your predictive analysis or anomaly detection task in big data analytics.

GITHUB REPOSITORY URL:

https://github.com/TamilSelvip-2004/Bigdata-analysis/settings/keys